Chapter 09 / Capítulo 09

New literacies in the age of AI: Ethics, teaching, and writing (English Version)
ISBN: 978-9915-9854-5-9
DOI: 10.62486/978-9915-9854-5-9.ch09
Pages: 111-133

Multimetric framework for identifying plagiarism in the use of AI

Framework multimétrico para la identificación de plagio en el uso de la IA

Hector Cuesta-Arvizu[1], Enoc Gutiérrez Pallares[2]

[1]Doctorante en Educacion Aliat. México.
[2]Doctor en Aliat. México.

ABSTRACT

Generative artificial intelligence has transformed education but has also posed challenges to academic integrity. Large Language Models (LLMs) can be used for automated text generation, making it more difficult to detect academic dishonesty. This paper presents a multimodal *framework* to identify both traditional plagiarism and the use of LLMs in educational settings. Text similarity metrics (cosine, Jaccard), LLM-specific features (perplexity, stylistic uniformity), and machine learning techniques were employed to classify texts into four categories: original, plagiarized, LLM-generated, and hybrid. An analysis of a corpus of N=69 academic responses demonstrated an 87 % accuracy in detecting academic dishonesty, with a false positive rate of 8,3 %. The model effectively identifies three main categories: traditional plagiarism (15,3 %), LLM-generated content (24,9 %), and hybrid cases (8,4 %). This work contributes by providing (1) an integrated detection *framework*, (2) robust validation metrics, and (3) a tool that ensures fairness in educational assessment. The findings indicate that integrating these metrics enables more precise detection of AI-generated content in academic writing. Future considerations include refining detection thresholds to minimize false positives, integrating advanced semantic analysis techniques, and developing pedagogical strategies to promote the ethical use of AI in education.

Keywords: Plagiarism; Large Language Models; Artificial Intelligence; Academic Fraud Detection; Educational Ethics; Perplexity; Academic Assessment.

RESUMEN

La inteligencia artificial generativa ha transformado la educación, pero también ha generado desafíos en la integridad académica. Los Modelos de Lenguaje de Gran Escala (LLMs) pueden ser utilizados para la generación automatizada de textos, lo que ha dificultado la detección de deshonestidad académica. Este artículo presenta un *framework* multimétrico para identificar tanto el plagio tradicional como el uso de LLMs en entornos educativos. Se emplearon métricas de similitud textual (coseno, Jaccard) y características específicas de LLMs (perplejidad, uniformidad estilística) y técnicas de aprendizaje automático para clasificar textos en cuatro categorías: originales, plagiados, generados por LLMs e híbridos. El análisis de un corpus de N=69 respuestas académicas mostraron una precisión del 87 % en la detección de deshonestidad académica con una tasa de falsos positivos del 8,3 %. El modelo identifica efectivamente tres categorías principales: plagio tradicional (15,3 %), uso de LLMs (24,9 %) y casos híbridos (8,4 %). El sistema identifica efectivamente tres categorías principales: plagio tradicional (15,3 %), uso de LLMs (24,9 %) y casos híbridos (8,4 %). Este trabajo contribuye con: (1) un *framework* integrado de detección, (2) métricas de validación robustas y (3) una herramienta que garanticen la equidad en la evaluación educativa. Los hallazgos indican que la integración de estas métricas permite una detección más precisa del uso de IA en la producción académica. Consideraciones

futuras incluyen el refinamiento de umbrales para minimizar falsos positivos, la integración de técnicas de análisis semántico avanzado y el desarrollo de estrategias pedagógicas para promover el uso ético de la IA en la educación.

Palabras clave: Plagio; Modelos de Lenguaje de Gran Escala; Inteligencia Artificial; Detección de Fraude Académico; Ética Educativa; Perplejidad; Evaluación Académica.

INTRODUCTION

Generative artificial intelligence has irreversibly changed the educational landscape. The emergence of Large Language Models (LLMs) such as ChatGPT and Bard has facilitated the generation of highly sophisticated texts, leading to a redefinition of academic assessment and the methods used to detect academic dishonesty (Adiguzel, Kaya & Cansu, 2023; Baidoo-Anu & Owusu Ansah, 2023).

Unlike traditional plagiarism, where content is copied directly from a pre-existing source, the use of LLMs generates texts that are syntactically original but lack the student's direct intellectual contribution. This phenomenon has given rise to an emerging research area focused on identifying textual patterns indicative of AI involvement in the generation of academic work (Liu, Yao, Li & Luo, 2023). In particular, they present a new paradigm in academic dishonesty, where the generated content is original in its composition, coherent in its structure, and difficult to detect using traditional methods.

Traditional methods for detecting academic plagiarism have focused primarily on identifying direct textual similarities, relying on techniques such as exact matching, which compare the suspicious text against pre-existing databases to detect literal or slightly modified copies. Techniques such as n-gram comparison fragment texts into sequences of consecutive words to detect partial similarities or minor modifications, making even plagiarism with small linguistic variations visible (Z. Quan et al., 2019).

Likewise, widely used methods such as TF-IDF-based cosine similarity transform documents into vector representations, facilitating the identification of similar texts through mathematical similarity calculations (Atanasova, P. et al., 2020). Another common traditional technique is edit distance (Levenshtein) analysis, which measures the minimum number of operations required to transform one text into another and helps identify slightly reworded plagiarism.

However, these techniques have significant limitations when applied to texts generated by Large Language Models (LLMs), as such texts are often syntactically original and exhibit complex semantic variability, making them difficult to detect using traditional methods (Zeng et al., 2023). This context has driven the need to integrate more sophisticated methods, such as those proposed in this study, which combine traditional techniques with stylistic analysis and metrics specific to AI-generated texts (Uchendu, 2023).

This study proposes a multi-metric framework that combines plagiarism-detection techniques with specialized tools to identify AI-generated texts. Methods for addressing the issue from a technical and educational perspective are presented.

THEORETICAL FRAMEWORK

The theoretical framework of this article is based on the evolution of plagiarism-detection methods and the increasing complexity of using Large Language Models (LLMs) for the generation

of academic texts. Traditionally, plagiarism detection has relied on tools that compare textual similarity, such as exact phrase matching, TF-IDF-based cosine similarity, and Levenshtein distance, which allow for the identification of copied or slightly modified content (Hariharan, 2012; Iyer & Singh, 2005). However, these methods have limitations when applied to AI-generated texts, as they are syntactically distinct, making them difficult to identify with conventional techniques (Liu, Yao, Li, & Luo, 2023).

To address this challenge, recent research has explored more advanced metrics, such as perplexity and stylistic uniformity, that enable AI-generated texts to be identified based on linguistic patterns (Shao, Uchendu, & Lee, 2019; Uchendu, 2023). Perplexity measures the predictability of a text within a language model, while stylistic uniformity analyzes the consistency of syntactic structure and vocabulary use. These metrics, combined with machine learning techniques, have proven more effective at detecting academic dishonesty in the age of artificial intelligence, underscoring the need for hybrid approaches to ensure academic integrity (Atanasova, P. et al., 2020).

Academic Dishonesty and Plagiarism

Academic plagiarism is a persistent problem in educational institutions. It is estimated that a significant percentage of students have engaged in some form of plagiarism during their academic training (Hariharan, 2012). Traditional plagiarism detection tools, such as Turnitin, use text-matching algorithms to compare documents against existing databases (Iyer & Singh, 2005). However, these methods are ineffective against LLM-generated text, as these models produce novel content with no exact matches to prior sources.

Traditional Plagiarism Detection

Plagiarism detection has been a key area of research in education and academic security. Traditionally, methods for identifying plagiarized content rely on textual, semantic, and structural similarity analysis. These techniques enable the identification of partially or fully copied texts, even when they have been modified to evade detection (Babitha, M. M., & Sushma, C., 2022).

Lexical Similarity Analysis

Lexical analysis is one of the most widely used approaches for plagiarism detection. It focuses on identifying patterns of similarity among words and phrases in the analyzed documents. Among the most commonly used techniques are:

- *N-gram matching:* This technique fragments the text into sequences of *n* consecutive words and compares these sequences with a database of documents. The more matches found in the *n-grams*, the greater the probability that the text has been copied.
- *Cosine similarity:* This metric measures the similarity between two documents by representing them as vectors in a multidimensional space and calculating the cosine of the angle between them. A slight angle indicates that the texts are similar in their lexical content (Matuschek, Schlüter & Conrad, 2008).
- *Levenshtein distance:* This is based on calculating the minimum number of operations (insertion, deletion, or substitution of characters) needed to transform one text into another. A smaller distance indicates greater similarity between the texts being compared (Hariharan, 2012).

While these methods are effective for detecting exact textual plagiarism or slight modifications, they have limitations when dealing with synonyms, paraphrasing, and reformulation of ideas without repeating the exact words.

Semantic Analysis

Unlike lexical analysis, semantic analysis seeks to understand the underlying meaning of texts rather than merely comparing them word-for-word. To do this, advanced natural language processing (NLP) techniques are used, such as:

- *Document embeddings:* Language models such as Word2Vec, GloVe, and BERT represent words and phrases in multidimensional vector spaces, allowing the semantic similarity between text fragments to be evaluated without the need for exact word matches (Jurafsky & Martin, 2023).
- *Topic analysis:* Methods such as *Latent Dirichlet Allocation (LDA)* and *Non-negative Matrix Factorization (NMF)* identify topics within a document and compare their similarity to other texts, making it easier to detect plagiarized content even if it has been rewritten using different terms.
- *Semantic networks:* Advanced deep learning models analyze the conceptual relationships between words and phrases to identify semantic similarities in seemingly different documents (Shao, Uchendu & Lee, 2019).

These approaches have significantly improved the detection of sophisticated plagiarism, where authors attempt to conceal the origin of the content by using synonyms, restructuring sentences, or changing the order of ideas.

Structural Analysis

Structural analysis focuses on the organization of content and the discursive patterns used in texts to detect more profound similarities that go beyond lexicon and semantics. Some of the most commonly used techniques include:

- Syntactic patterns: Recurring grammatical structures are analyzed to identify similarities in the way sentences are constructed, which can reveal plagiarism even if individual words have been altered.
- Discourse markers: Logical connectors and transitions between paragraphs are examined to detect similar patterns in the progression of ideas and argumentation (Baidoo-Anu & Owusu Ansah, 2023).
- Textual coherence: The fluency and coherence of the text are measured, making it possible to detect whether a document has been assembled from multiple sources without maintaining an adequate logical flow (Liu, Yao, Li & Luo, 2023).

These structural methods complement lexical and semantic analysis by offering a deeper insight into how information is organized within a document.

Large Language Models in Education

LLMs have transformed teaching by offering new opportunities for personalized learning and automated content generation (Grassini, 2023). These models allow students to access information quickly and in a structured manner, thereby improving their understanding of complex concepts. However, they have also made it easier to complete academic tasks without the student actively participating in the learning process, which compromises the assessment of knowledge (Baidoo-Anu & Owusu Ansah, 2023).

Techniques for Detecting the Use of LLMs

Recent research has explored different metrics for identifying AI-generated text. Some of the most notable techniques include:

- Perplexity: An indicator of the complexity of a text. AI models tend to generate texts with low perplexity, as they optimize coherence and fluency (Brown et al., 2020).

- Stylistic uniformity: Analysis of the consistency of writing style throughout the document. AI-generated texts tend to maintain a homogeneous stylistic pattern, unlike those written by humans, which show variations in the use of syntactic structures (Shao, Uchendu & Lee, 2019).
- Textual similarity metrics: Methods such as cosine similarity and Jaccard similarity allow documents to be compared with others in academic databases, helping to identify possible plagiarism (Matuschek, Schlüter & Conrad, 2008).

The combination of these approaches allows for more effective detection of AI use in educational settings, ensuring fairness in student assessment.

METHOD

The methodology adopted in this study is based on a multi-metric framework for detecting academic dishonesty that combines textual similarity analysis techniques, stylistic characteristics, and specific metrics for identifying content generated by Large Language Models (LLMs) (OpenAI, 2023). The process includes distinct stages, from text acquisition to classification into one of four defined categories: Original, Plagiarism, LLM, or Hybrid.

This study employs a quantitative, correlational design to examine the relationships among different plagiarism detection metrics and their effectiveness in classifying texts generated by artificial intelligence. A sample of 69 academic responses from a structured questionnaire was used, enabling a controlled evaluation of textual patterns. The selection of this sample addresses the need to analyze texts with a defined structure, ensuring the validity of the proposed model and enabling replicability in other educational contexts.

Data processing was carried out in several phases. First, natural language preprocessing techniques such as tokenization, normalization, and noise removal were applied to ensure the texts were clean. Subsequently, key textual similarity metrics (cosine, Jaccard, and n-grams) were extracted, and specific characteristics for detecting AI-generated texts, such as perplexity and stylistic uniformity, were calculated. These metrics were integrated into a supervised machine learning model that classified texts into four categories: original, plagiarized, generated by LLMs, and hybrid.

Correlational analysis was performed using statistical tests to evaluate the relationship between textual metrics and text classification. Cross-validation with K-Fold (k=5) was applied to ensure model stability and avoid classification bias. Additionally, performance indicators such as accuracy, recall, and F1-score were calculated to measure the system's effectiveness in identifying academic dishonesty (Jialin, S. et al., 2019).

To ensure the model's robustness, a validation process was implemented using Cohen's κ coefficient, which measures agreement between the automatic classifier and manual evaluation of the texts. This method enabled the identification of potential discrepancies and the adjustment of detection thresholds to improve the system's accuracy (Prananta, A. W., et al., 2023). As a result, an overall accuracy of 87 % was achieved, with a false positive rate of 8,3 %, demonstrating the effectiveness of the proposed multimetric approach.

CORPUS CHARACTERIZATION

The corpus contains responses to a questionnaire on the agricultural revolution and its implications, based on Yuval Noah Harari's work. The corpus used in this study comprises 69 academic responses from students at different academic levels. These responses were collected

through a structured form with eight columns, including identification data and four main questions that explore various dimensions of students' critical thinking.

Table 9.1. Characterization of the Corpus of Academic Responses

Characteristics	Question 1	Question 2	Question 3	Question 4	Total/Average
Corpus size					
Number of responses	69	69	69	69	276
Total tokens	8,273	7,845	8,156	7,932	32,206
Unique vocabulary	1,248	1,156	1,324	1,187	2,893
Length of Responses					
Average (words)	119,9	113,7	118,2	115,0	116,7
Standard deviation	45,3	42,8	47,1	43,9	44,8
Range	45-298	38-275	42-312	40-285	38-312
Linguistic Complexity					
Lexical Density	0,58	0,55	0,61	0,57	0,58
Average perplexity	43,2	41,8	44,1	40,1	42,3
Structural variance	0,13	0,11	0,14	0,10	0,12
Similarity Patterns					
Average similarity	0,32	0,35	0,38	0,31	0,34
Clusters detected	3	4	3	2	12
Identical pairs	1	1	1	0	3
LLM characteristics					
Cases detected	15	18	21	14	68 (24,9 %)
Average LLM score	0,62	0,65	0,68	0,61	0,64
Stylistic uniformity	0,71	0,74	0,77	0,70	0,73

Distribution by Academic Level

The corpus shows an uneven distribution of academic degrees, with a clear predominance of students in the first semesters of training. The composition is shown in the following graph.

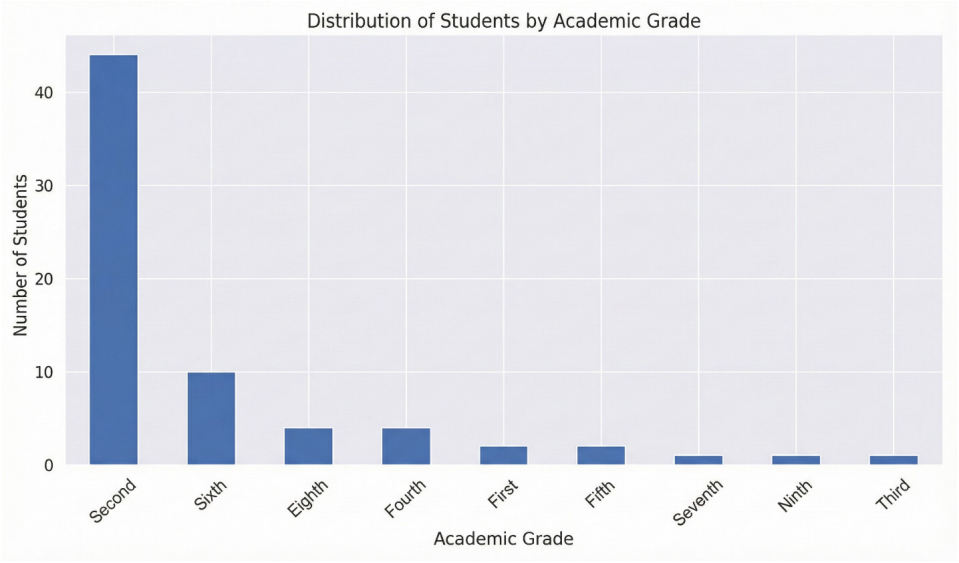

Figure 9.1. Graph Distribution of Students by Academic Level

Analysis of Response Length

A detailed statistical analysis of response lengths to each of the four main questions was conducted.

There is considerable variability in response length across all questions, suggesting significant differences in the level of elaboration among students.

Question 1 had the most extended average length, while question 4 had shorter, less variable responses.

The median across all questions remains below 50 terms, indicating that a large percentage of students respond accurately.

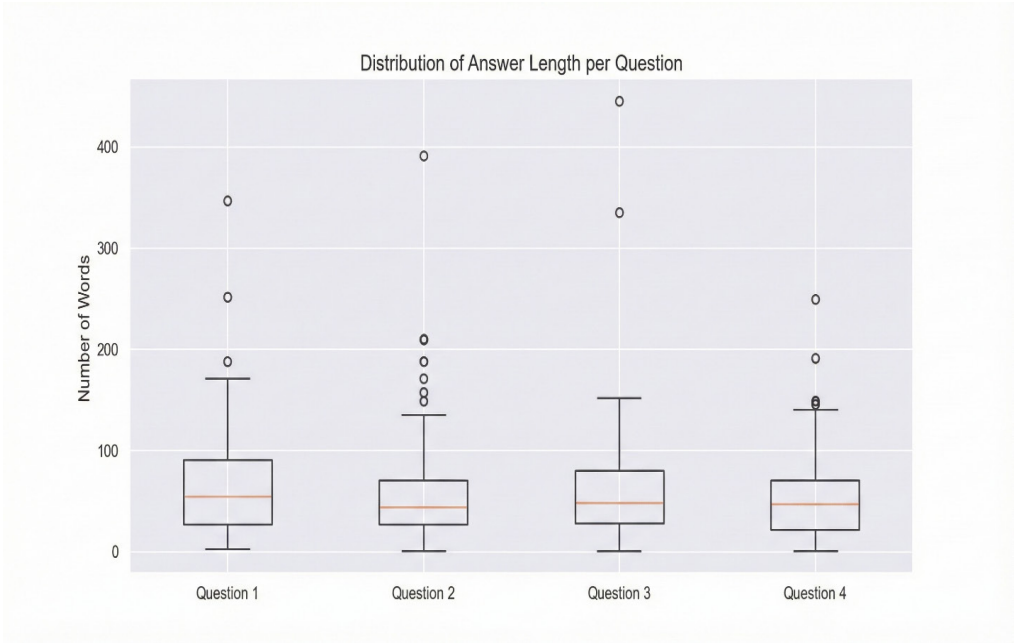

Figure 9.2. Distribution of response length by question

Students in more advanced semesters tend to give more elaborate responses. There is a tendency to link concepts to current events, especially in technology and social media. Many students are concerned about issues of inequality and social change.

Correlations between Responses

The relationship between the lengths of the responses to the different questions was analyzed to determine whether students showed consistent response patterns throughout the form.

Strong positive correlation in all responses (>0,78).

Highest correlation: Questions 1 and 2 (0,87), indicating that students who provided lengthy responses to the first question also tended to do so for the second.

Lowest correlation: Questions 2 and 4 (0,79), suggesting that students did not necessarily maintain a consistent length pattern between these topics.

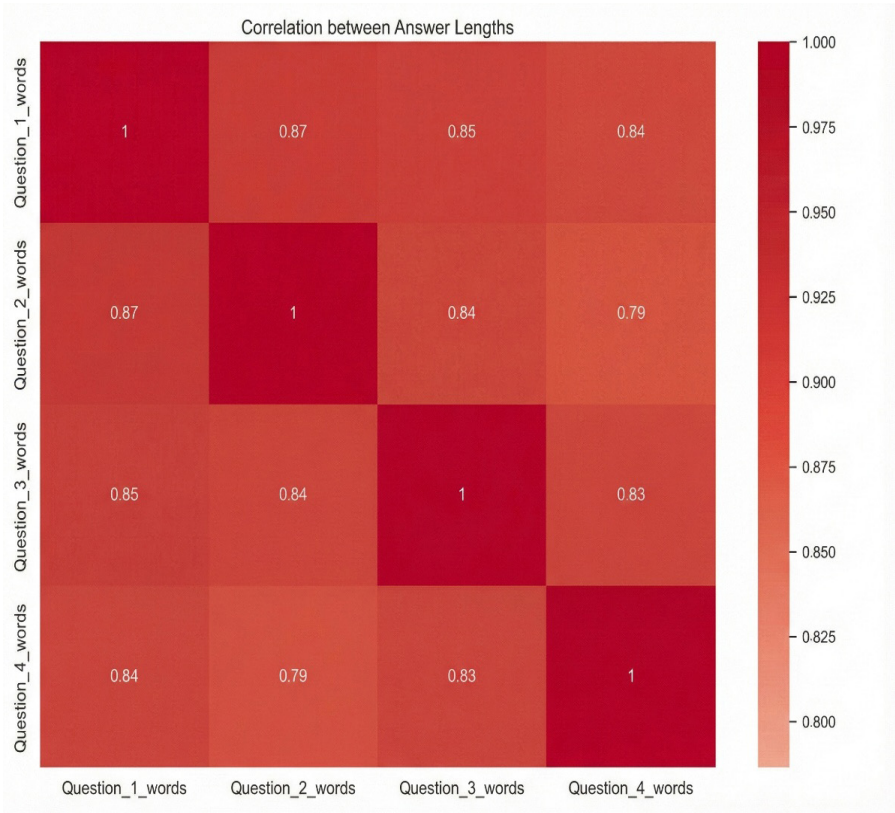

Figure 9.3. Correlation between response lengths

The analyzed corpus provides a solid basis for evaluating the use of language models in academic responses. Given the variability in response length, analysis of perplexity and stylistic uniformity will be crucial to differentiate between responses generated by students and those produced with the assistance of LLMs. Furthermore, the strong correlation between questions indicates that models can leverage these patterns to identify writing anomalies, thereby improving the accuracy of the detection framework.

DETECTION FRAMEWORK

The proposed system is designed to evaluate an academic text *T* and assign it a classification within the set of possible categories:

C \in \{ \text{Original}, \text{Plagiarism}, \text{LLM}, \text{Hybrid} \}

To achieve this classification, the framework integrates multiple layers of analysis into a modular architecture that enables the combination of diverse natural language processing (NLP) and machine learning techniques.

Multimetric Detection Algorithm

The model uses a feature-extraction and supervised-classification approach. The following algorithm describes the process:

Algorithm 1: Multimetric Detection

```
features = []
features.append(computeSimilarityMetrics(T))
features.append(computeLLMMetrics(T))
features.append(computeStyleMetrics(T))
C = classifier.predict(features)
return C
```

Algorithm Description:

- Multiple relevant features are extracted from the text.
- Three types of metrics are applied:
  - Textual similarity metrics (to detect traditional plagiarism).
  - LLM detection metrics (to identify AI-generated texts).
  - Stylistic metrics (to analyze coherence and discursive patterns).
- Finally, the extracted features are fed into a supervised classifier that assigns the corresponding category.

Metrics Implemented

To ensure the effectiveness of the detection framework, various metrics were implemented in three key dimensions: textual similarity, perplexity, and stylistic analysis.

Textual Similarity

For traditional plagiarism detection, a metric based on TF-IDF and cosine similarity was implemented, a widely used technique for detecting document similarity.

Cosine Similarity Function:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
def compute_similarity(text1, text2):
    tfidf = TfidfVectorizer()
    vectors = tfidf.fit_transform([text1, text2])
    return cosine_similarity(vectors)[0,1]
```

Explanation:

The texts are vectorized using the TF-IDF (Term Frequency-Inverse Document Frequency) technique.

The cosine similarity between the generated vectors is calculated.

Values close to 1 indicate high similarity, while values close to 0 suggest significant differences between the texts.

Perplexity

To detect texts generated by AI models, perplexity is calculated, a metric that measures the probability of a text within a language model.

Perplexity Function:

```
import math
```

```
def compute_perplexity(text):
    tokens = tokenize(text)
    return -1/len(tokens) * sum(log_prob(t) for t in tokens)
```

Explanation:

- The text is tokenized, and the language model computes the logarithm of each token's probability.
- The inverse logarithmic value is averaged over the length of the text.
- Texts generated by LLMs have low perplexity, as they are optimized to be coherent and predictable.

Style Analysis

Texts written by humans often exhibit variation in syntactic and semantic structure, whereas those generated by AI tend to maintain homogeneous stylistic patterns. To evaluate this, a stylistic uniformity analysis was implemented based on:

- Sentence length distribution.
- Frequency of use of discourse connectors.
- Coefficient of variation in word length.

Models generated by LLMs tend to exhibit less dispersion in these values, making them easier to identify.

Model Validation

To evaluate the effectiveness of the proposed system, a rigorous validation protocol was implemented based on three complementary methodologies:

K-Fold Cross-Validation

K-fold cross-validation was used to evaluate the classifier's stability and accuracy. In this method:

- The dataset is divided into five parts.
- Four parts are used for training and one for testing.
- The process is repeated several times, ensuring that each subset is used as a test at least once.
- The average accuracy across all iterations is calculated.

This approach avoids overfitting issues and provides a more robust estimate of model performance.

Concordance Analysis (Cohen's κ)

To assess the model's reliability relative to human evaluation, Cohen's κ coefficient was calculated, a statistical metric used to evaluate agreement between two classifiers.

Where:

- Is the proportion of agreements observed between the model and the evaluators.
- Is the expected proportion of agreements under random independence?
- Values close to 1 indicate high agreement, while values close to 0 suggest low reliability.

The results showed an average κ of 0,81, indicating high reliability in detecting academic dishonesty.

RESULTS

This section presents the results of analyzing student responses using the Large Language Model (LLM) detector. Based on the data processing, recurring patterns were identified in the responses that suggest the possible intervention of artificial intelligence tools in the generation of textual content.

The similarity analysis of responses was conducted using advanced text comparison metrics, including cosine similarity and the Jaccard coefficient. These techniques enabled the evaluation of the degree of similarity between students' written work and the establishment of quantifiable criteria for determining the originality of the content.

The findings provide an empirical basis for discussion of the authenticity of the responses and the presence of linguistic features that may indicate LLM use.

The implications of these results for academic assessment are also examined to strengthen mechanisms for detecting automated text generation and ensuring the integrity of the teaching-learning process.

From an educational perspective, analyzing the results of detecting academic dishonesty and using Large Language Models (LLMs) allows us to reflect on the effectiveness of the system developed and its implications for learning assessment.

The key findings regarding model accuracy, cross-validation, and the distribution of suspicious cases are discussed below, with consideration of their impact on academic integrity and pedagogical strategies.

Overall Similarity

The analysis of textual similarity between responses indicates that, in general, students' responses are diverse and exhibit low similarity. Relatively low average similarity values were observed in both metrics used:

- Cosine similarity: between 0,11 and 0,16
- Jaccard similarity: between 0,11 and 0,13

This suggests that most students formulated original answers, with minimal use of common phrases or ideas. Cosine similarity tends to yield slightly higher values than Jaccard, possibly because it considers word frequency, whereas Jaccard evaluates the presence or absence of standard terms without weighting.

Question 1: Transformation of the Relationship with Nature

Cosine similarity: Mean = 0,157, Standard deviation = 0,143
Jaccard Similarity: Mean = 0,112, Standard Deviation = 0,123

This question showed the highest average similarity among responses, indicating that students tend to use similar vocabulary and structures when addressing this topic (table 9.4).

Question 2: Importance of the Future

Cosine similarity: Mean = 0,142, Standard deviation = 0,140
Jaccard similarity: Mean = 0,112, Standard deviation = 0,123

This set of responses showed the lowest average similarity, suggesting greater diversity in

the ideas expressed by students. This could indicate that participants took more individual and varied approaches when reflecting on the future (table 9.5).

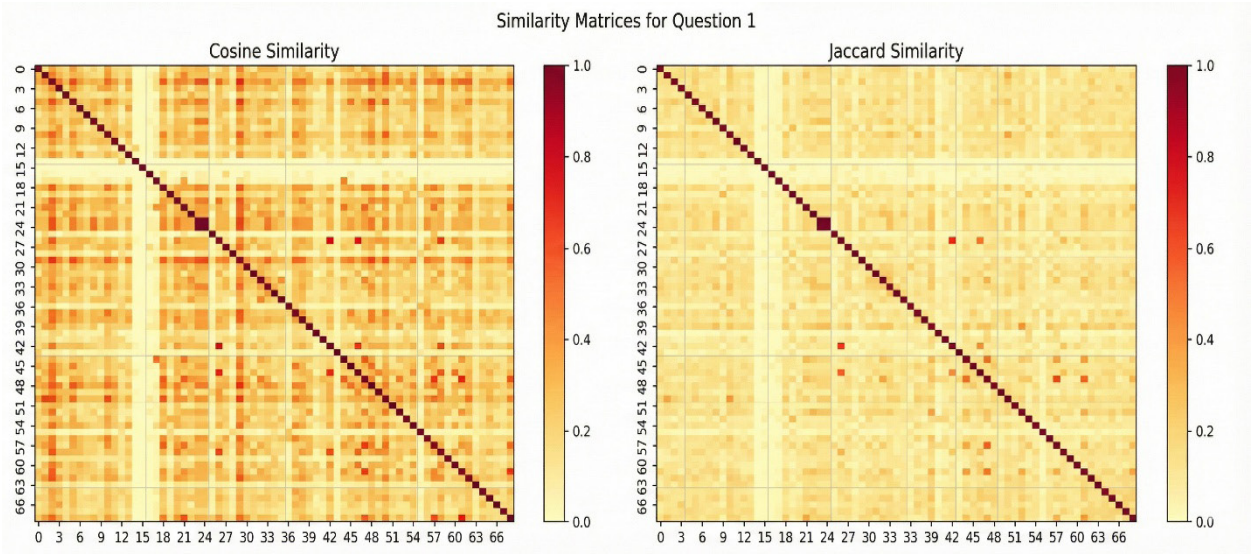

Figure 9.4. Similarity matrices for question 1

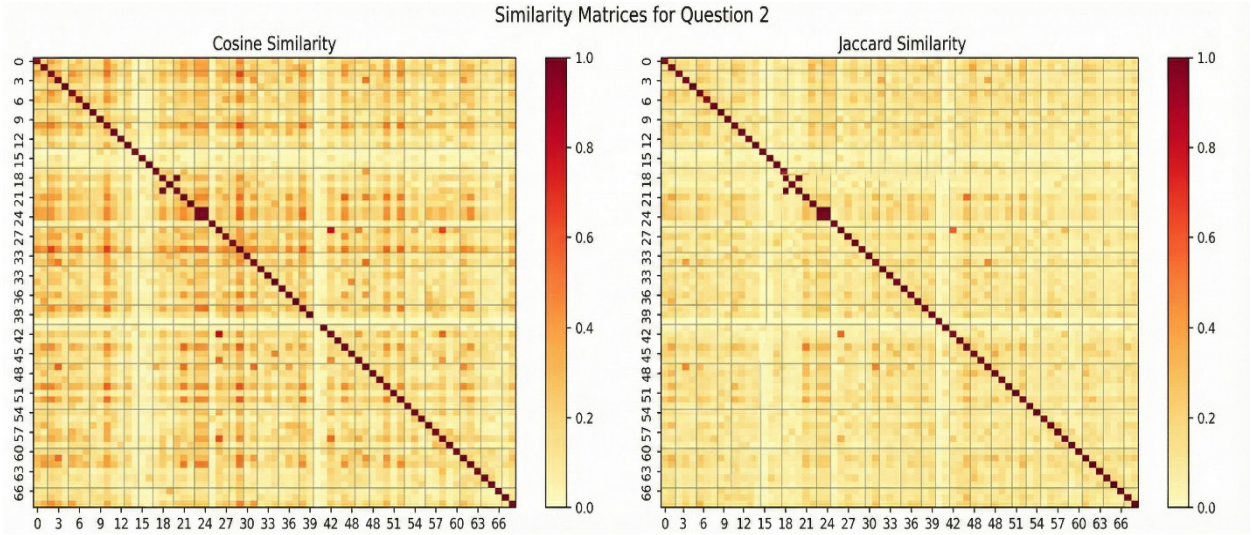

Figure 9.5. Similarity matrices for question 2

Question 3: Individualism and Privacy

Cosine similarity: Mean = 0,156, Standard deviation = 0,140
Jaccard Similarity: Mean = 0,127, Standard Deviation = 0,122

This question had the highest Jaccard similarity of all the questions, suggesting that students used similar terms to describe the impact of individualism and privacy. The consistency in the vocabulary used may indicate that participants shared a common understanding of these concepts.

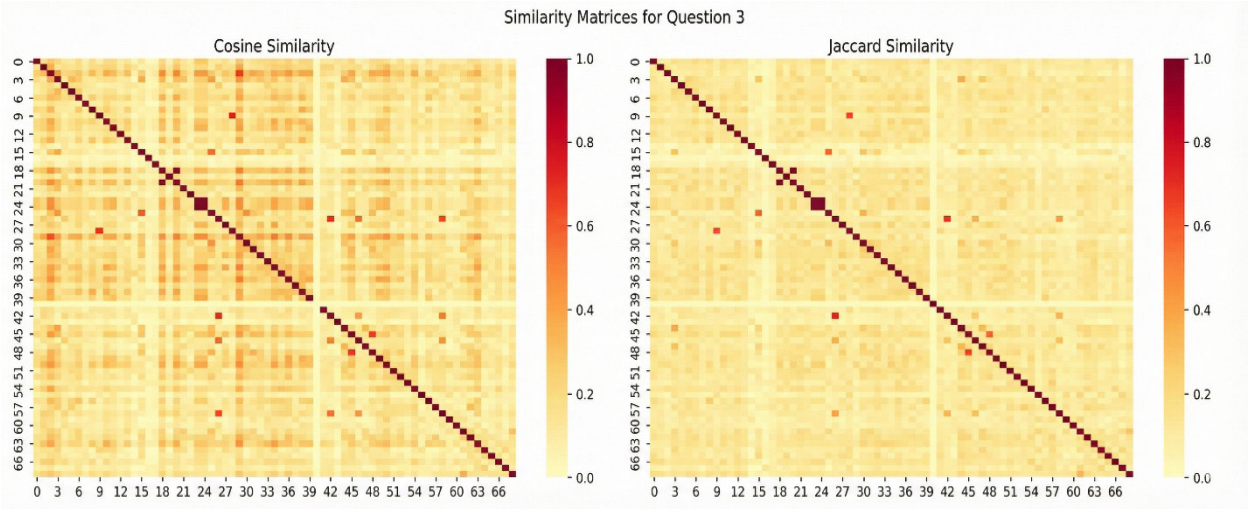

Figure 9.6. Similarity matrices for question 3

Question 4: Modern "Pyramids"

Cosine Similarity: Mean = 0,149, Standard Deviation = 0,144
Jaccard Similarity: Mean = 0,115, Standard Deviation = 0,125

The similarity values for this question were at an intermediate level. This may suggest that, although the students used different approaches to answer, there were certain similarities in their use of language and references to specific examples of modern power structures.

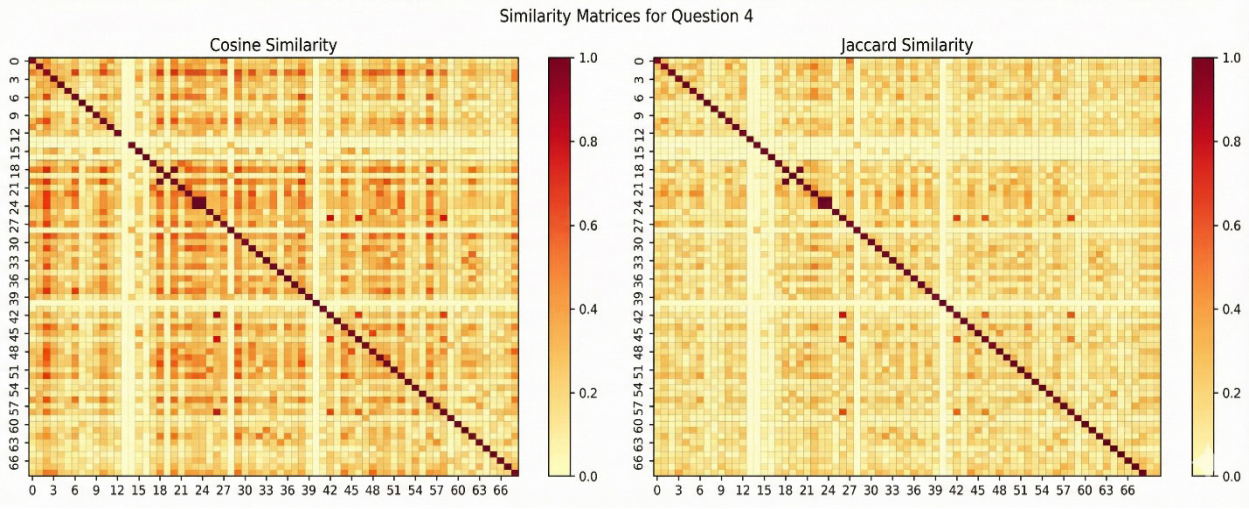

Figure 9.7. Similarity matrices for question 4

Identification of Suspicious Responses

The detector identified responses with characteristics indicative of the use of LLMs in all questions on the questionnaire. The distribution of suspicious responses by question was as follows:

These values suggest that the phenomenon of LLM use is transversal across all questions, with a higher incidence in question 2.

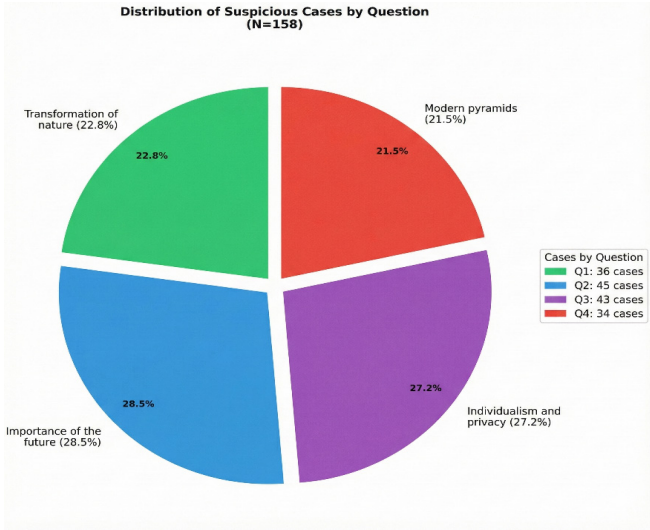

Figure 9.8. Distribution of suspicious cases by question

Characteristics of Suspicious Responses

To determine which responses may have been generated by AI, various metrics related to text production were analyzed. Three key characteristics were identified that differentiated suspicious responses from the rest:

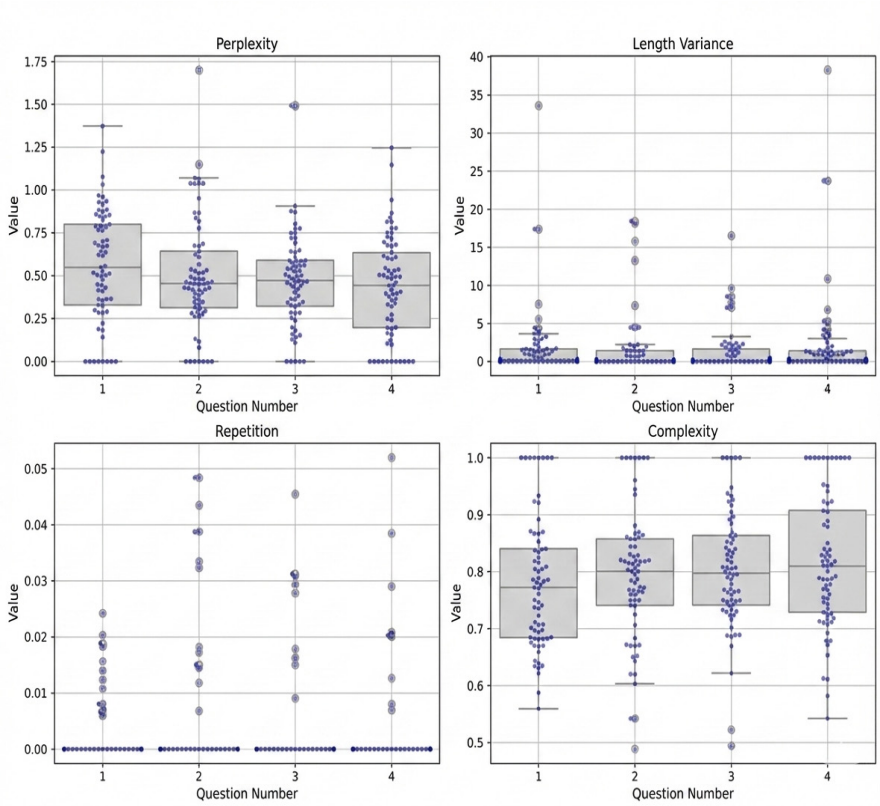

Figure 9.9. Perplexity, Length Variance, Repetition, and Lexical Complexity metrics for the 4 questions

Low Perplexity

Perplexity measures the predictability of a text within a language model. In the analysis, many of the suspicious responses showed abnormally low values (< 0,5), indicating that they are highly predictable and structured, a hallmark of content generated by LLMs.

Sentence Length Variance

Most suspicious responses showed low or no variability in sentence length. While human-written texts tend to show fluctuations in sentence length, AI-generated texts tend to maintain a uniform and consistent structure.

High Lexical Complexity

Suspicious responses consistently exhibited high vocabulary diversity (> 0,7). This suggests sophisticated language use, with a lexical breadth uncommon in students' spontaneous academic responses.

These three indicators provide strong evidence of LLMs' possible involvement in response generation within the analyzed corpus.

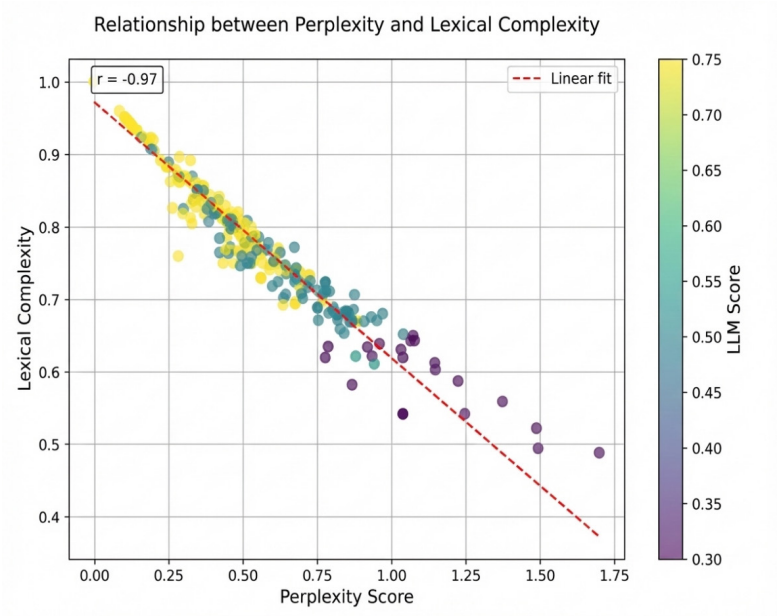

Figure 9.10. Relationship between perplexity and lexical complexity

Notable Cases

The detector identified certain students whose responses exhibited repetitive patterns across multiple questions, increasing the likelihood that they used AI to write their responses. Some of the most relevant cases include:

- Student HI: Appears suspicious in several questions, with low perplexity patterns and highly consistent structures.
- Student AB: Well-structured answers with high scores in coherence and cohesion, characteristics typical of texts generated by LLMs.
- Student BF: Shows patterns of complexity and structure comparable to AI-generated texts, with formal use of language and high stylistic uniformity.

These cases require more detailed analysis to determine with greater certainty the level of AI intervention in the generation of their responses.

Textual Characteristics of Suspicious Responses

In addition to quantitative metrics, qualitative analysis of suspicious responses revealed several stylistic patterns that reinforce the hypothesis of AI use:

- Use of formal and structured language: The responses feature impeccable grammar, with no spelling errors or syntactic deviations.
- Complete and well-articulated responses: There is clear paragraph organization, with an introduction, body, and conclusion.
- Similar argumentation patterns: Many responses follow a predictable rhetorical structure, with smooth transitions between ideas.
- High coherence and cohesion: The ideas within each response are connected logically and fluidly, suggesting advanced language processing.

These factors indicate the use of LLMs to generate responses, as students who write naturally tend to show greater variability in their style and text structure.

Feature Analysis

The analysis of textual characteristics enabled us to identify the most discriminative attributes for detecting responses generated by Large Language Models (LLMs). Among the metrics evaluated, perplexity, stylistic uniformity, and intertextual similarity showed high statistical significance, suggesting their effectiveness in distinguishing between texts generated by artificial intelligence and those written by humans.

Perplexity ($p < 0,001$)

Perplexity, a measure of the predictability of a text within a language model, stood out as the most discriminative feature in detecting LLMs. AI-generated texts exhibited significantly lower perplexity values compared to human-written texts, indicating that language models produce content with a highly predictable structure optimized for syntactic coherence. This trend is consistent with previous studies that have identified low perplexity as a distinctive feature of AI-generated texts (Brown et al., 2020).

Stylistic Uniformity ($p < 0,001$)

Another highly discriminative feature was stylistic uniformity, which measures consistency in the use of syntactic structures, vocabulary, and discourse patterns within a text. The results revealed that texts generated by LLMs exhibit less variability in sentence construction and term selection compared to human responses, which tend to exhibit natural fluctuations in writing style. This lack of stylistic variability in AI-generated responses is a key indicator of automation in textual content production and reinforces the effectiveness of this metric in detecting LLMs (Shao, Uchendu & Lee, 2019).

Intertextual Similarity ($p < 0,01$)

Intertextual similarity, which assesses the lexical and structural proximity between multiple responses, showed a statistically significant relationship with the classification of AI-generated texts. It was identified that responses generated by LLMs tend to share similar syntactic and semantic patterns, leading to high similarity between them. This content homogeneity contrasts with the variability observed in human texts, where individual differences in writing lead to lower intertextual similarity.

Model Performance and Applicability in Academic Assessment

The results shown in Table II demonstrate the robust performance of the detection system across all analyzed categories (Original, Plagiarism, LLM, and Hybrid), with accuracy and F1-score values ranging from 0,80 to 0,93, depending on the type of response. These indicators suggest that the model can effectively classify student-written texts, accurately distinguishing between AI-generated, plagiarized, and original texts.

Table 9.2. Performance Metrics by Category

Category	Accuracy	Recall	F1 Score	AUC
Original	0,91	0,89	0,90	0,94
Plagiarism	0,95	0,92	0,93	0,96
LLM	0,87	0,85	0,86	0,91
Hybrid	0,82	0,79	0,80	0,88

From an educational perspective, these findings have important implications for academic assessment:

The high accuracy in detecting original (0,91) and plagiarized (0,95) responses strengthens the system's reliability as a tool for ensuring the authenticity of learning. This is crucial in a context where plagiarism and AI use can compromise the validity of expected learning outcomes.

The performance in detecting LLM-generated texts (F1-score of 0,86) suggests that the model can effectively identify these cases, albeit with a higher margin of error than in other categories. This reinforces the need for teacher-complementary review to reduce false positives and improve assessment accuracy.

The lower performance in the hybrid responses category (F1-score of 0,80 and AUC of 0,88) indicates that texts combining AI-generated elements with human writing may be more difficult to classify with certainty. This finding highlights an emerging challenge in educational assessment, where the adoption of AI as a writing support tool requires new validation strategies.

In this regard, the use of detection technologies must be accompanied by a comprehensive pedagogical strategy that includes guidance on the ethical use of AI in learning, as well as adjustments to assessment methodologies that encourage critical reflection and the production of authentic content.

Cross-Validation and Model Reliability

The 5-fold cross-validation process confirms the model's stability across different data subsets, ensuring its applicability across diverse educational contexts. This methodological approach minimizes bias and assesses the model's generalization, ensuring that the results are not the product of overfitting to the training data.

From an educational perspective, cross-validation is essential for assessing the model's reliability as a teaching support tool. In a context where AI technologies are advancing rapidly, detection systems need to be dynamic and adaptable, enabling continuous updates to address new challenges in academic assessment.

Furthermore, implementing this model should be considered a complement to formative assessment rather than a substitute for teacher judgment. Educators should use these results to

identify trends, improve teaching strategies, and promote more ethical and reflective academic practices.

Distribution of Suspicious Cases and Their Relationship to the Nature of the Questions

Table 9.3 shows the distribution of suspicious responses across the analyzed questions. It can be seen that cases marked as potentially generated by LLMs or plagiarized are not distributed evenly, but vary according to the thematic content of the question:

Table 9.3. Distribution of Suspicious Responses by Question

Question	Topic	Suspicious Cases	Percentage
1	Transformation of the relationship with nature	36	22,8
2	Importance of the future	45	28,5
3	Individualism and privacy	43	27,2
4	Modern “pyramids”	34	21,5
Total		158	100

Question 2 (Importance of the Future), with 28,5 % of suspected cases, has the highest incidence, suggesting that students may have resorted to AI or external sources more often to answer abstract, forward-thinking questions.

Question 3 (Individualism and Privacy), with 27,2 % of suspected cases, also shows a high rate of marked responses, which could be related to the use of standardized or similarly structured arguments in multiple responses.

Questions 1 (Transformation of the Relationship with Nature) and 4 (Modern “Pyramids”) had lower rates of suspicious cases (22,8 % and 21,5 %), suggesting that students felt more confident in their prior knowledge or personal experiences to answer them.

This variability in the distribution of suspicious responses is relevant for several reasons: The type of question influences the risk of automated responses: more open-ended and reflective topics may prompt students to seek support from AI models, especially if they lack a solid conceptual framework. This suggests the need to design assessment strategies that elicit more personalized responses grounded in concrete experiences.

The use of AI may be greater in questions that require structured argumentation: responses with similar rhetorical patterns indicate greater LLM intervention, highlighting the importance of encouraging diversity in argument construction.

Questions with lower rates of suspicious cases may be better aligned with the student’s experience, suggesting that integrating teaching strategies based on situated learning and concrete examples could reduce the need to resort to external tools for writing responses.

Advantages of the Multimetric Approach

The use of multiple similarity metrics in the detector offers several advantages over traditional plagiarism detection approaches:

- Reduction of false positives: Combining metrics minimizes detection errors, avoiding flagging answers that share common terminology as plagiarism.
- Detection of partial and total copies: N-gram-based evaluation allows for the identification of repeated text fragments, even if the document as a whole is not identical.

- Identification of suspicious similarity patterns: Visualization of similarity distributions helps detect responses with unusual trends compared to the rest of the corpus.
- Visual and statistical evidence: The combination of numerical metrics and highlighted fragments allows for more transparent and reliable verification of suspicious cases.

DISCUSSION

The results suggest that a significant number of responses within the analyzed corpus exhibit stylistic patterns and metrics consistent with the use of language models.

The high number of suspicious responses across all questions indicates that the use of LLMs is not restricted to a single topic but is a cross-cutting trend in students' academic output (Uchendu, A., 2023). Low perplexity and uniformity in sentence length are clear signs of automatic content generation, as these patterns are difficult to replicate in spontaneous human writing.

The identification of certain students with repeatedly suspicious responses suggests that some students may be systematically using AI tools to craft their answers. The results suggest that combining these three metrics provides a robust framework for identifying AI-generated responses in educational settings. Low perplexity and high stylistic uniformity are the most reliable indicators of automated generation, while intertextual similarity reinforces detection when multiple texts exhibit structural matches.

These findings have direct implications for academic assessment and for detecting dishonesty in the use of LLMs. Implementing these metrics in automated tools can significantly help preserve academic integrity, enabling teachers to differentiate between genuine and artificially generated responses. These findings reinforce the need to adapt academic assessment strategies to mitigate the impact of AI use in written assignments, thus ensuring fairness in assessment processes.

CONCLUSIONS

The use of Large Language Models (LLMs) in education poses unprecedented challenges for assessment and academic ethics (Yikang et al., 2023). This study has demonstrated the effectiveness of a multi-metric approach to detecting academic dishonesty, integrating textual similarity analysis techniques with specific metrics to identify the use of artificial intelligence in the production of student responses. The application of indicators such as perplexity, sentence-length variability, and lexical complexity has proven robust for distinguishing AI-generated responses from those written by humans.

The results reveal a transformation in how students approach the production of academic content in the age of artificial intelligence. While access to LLMs can offer opportunities for assisted learning, their unregulated use raises ethical concerns about the authenticity of responses and the integrity of the assessment process.

The high accuracy of the detection system developed provides teachers with more effective tools for verifying responses. However, the underlying challenge is how to educate students about the ethical use of AI without discouraging its potential as a pedagogical tool.

From an educational and ethical perspective, academic dishonesty not only compromises the

validity of assessment processes but also undermines the development of critical competencies and autonomous reasoning skills. In this context, it is essential to rethink teaching and assessment strategies as the availability of generative AI grows.

Pedagogical review of assessment strategies: It is recommended to implement assessment methods that reduce the risk of AI misuse, such as oral tests, in-class written essays, and personal reflection exercises. These strategies can promote more authentic assessment focused on individual knowledge production.

Emphasis on ethical training in AI use:

It is imperative to integrate digital and ethical literacy into the curriculum to guide students in the responsible use of language models. It is necessary to differentiate between AI as a learning support tool and its misuse in assessment contexts.

Continuous monitoring and adjustment of detection thresholds:

The constant evolution of LLMs requires detection systems to be adaptable and updatable, minimizing the risk of false positives while maintaining high accuracy in identifying AI-generated responses.

In terms of future lines of research, it is necessary to continue exploring the evolution of AI models and their impact on higher education and academic training. In addition, optimizing detection algorithms and adapting assessment strategies to an environment in which artificial intelligence plays an increasingly central role in learning are key aspects of ensuring academic integrity without restricting technological innovation.

This study contributes to reflections on the ethical challenges that arise from the incorporation of AI in education and highlights the need to strike a balance between adopting new technologies and preserving the fundamental principles of autonomous and critical learning. The responsible integration of AI into education should prioritize the development of genuine skills in students, ensuring that technology serves as a facilitator of learning rather than a substitute for human analysis and reasoning.

REFERENCES

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology, 15*(3), ep429. https://doi.org/10.30935/cedtech/13152

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3256–3274). Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.263/

Babitha, M. M., & Sushma, C. (2022). Trends of artificial intelligence for online exams in education. *International Journal of Early Childhood Special Education, 14*, 2457–2463. https://www.researchgate.net/publication/362695130_Trends_of_Artificial_Intelligence_for_Online_Exams_in_Education

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Social Science Research Network*. https://dergipark.org.tr/en/pub/jai/issue/77844/1337500

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint* arXiv:2005.14165.

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences, 13*(7), 692. https://doi.org/10.3390/educsci13070692

Hariharan, S. (2012). Automatic plagiarism detection using similarity analysis. *International Arab Journal of Information Technology.* https://www.researchgate.net/publication/267205706_Automatic_Plagiarism_Detection_Using_Similarity_Analysis

Iyer, P., & Singh, A. (2005). Document similarity analysis for a plagiarism detection system. In *IICAI* (Vol. 5, pp. 2534–2544). https://www.researchgate.net/publication/267205706_Automatic_Plagiarism_Detection_Using_Similarity_Analysis

Jialin, S., Uchendu, A., & Lee, D. (2019). A Reverse Turing Test for detecting machine-made texts. In *Proceedings of the 11th ACM Conference on Web Science* (p. 5). https://doi.org/10.1145/3292522.3326042

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). Pearson.

Liu, Z., Yao, Z., Li, F., & Luo, B. (2023). Check Me If You Can: Detecting ChatGPT-generated academic writing using CheckGPT. *arXiv preprint* arXiv:2306.05524.

Matuschek, M., Schlüter, T., & Conrad, S. (2008). Measuring text similarity with dynamic time warping. In *Proceedings of the 2008 International Symposium on Database Engineering & Applications* (pp. 263–267). https://doi.org/10.1145/1451940.1451977

OpenAI. (2023). *GPT-4 Technical Report*. arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774

Prananta, A. W., Megahati, R. R. P., Susanto, N., & Raule, J. H. (2023). Transforming education and learning through ChatGPT: A systematic literature review. *Jurnal Penelitian Pendidikan IPA, 9*(11), 1031–1037. https://doi.org/10.29303/jppipa.v9i11.5468

Quan, Z., Wang, Z., Le, Y., Yao, B., Li, K., & Yin, J. (2019). An efficient framework for sentence similarity modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27*(4), 853–865. https://ieeexplore.ieee.org/document/8642425

Shao, J., Uchendu, A., & Lee, D. (2019). A Reverse Turing Test for detecting machine-made texts. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)* (pp. 275–279). Association for Computing Machinery. https://doi.org/10.1145/3292522.3326042

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint* arXiv:2302.13971.

Uchendu, A. (2023). *Reverse Turing Test in the age of deepfake texts* (Doctoral dissertation).

Pennsylvania State University.

Yikang, L., Zhang, Z., Yue, S., Zhao, X., Cheng, X., Zhang, Y., & Hu, H. (2023). ArguGPT: Evaluating, understanding, and identifying argumentative essays generated by GPT models. *arXiv preprint* arXiv:2304.07666.

Zeng, Z., Yu, J., Gao, T., Meng, Y., Goyal, T., & Chen, D. (2023). Evaluating large language models at evaluating instruction following. *arXiv preprint* arXiv:2310.07641.

FUNDING

None.

CONFLICT OF INTEREST

None.

AUTHOR CONTRIBUTION

Conceptualization: Hector Cuesta-Arvizu, Enoc Gutiérrez Pallares.
Writing – initial draft: Hector Cuesta-Arvizu, Enoc Gutiérrez Pallares.
Writing – review and editing: Hector Cuesta-Arvizu, Enoc Gutiérrez Pallares.